Meta-moments in meta-coherence; fieldnotes on falsification

Benjamin James

June 27, 2026

The standard worry about a theory is that it might be false. The more revealing worry is that it might be unkillable. A theory that cannot be wrong is not thereby strong; it has merely arranged its commitments so that no possible observation counts against them, and that arrangement is usually invisible to the theory that made it. Popper called the line between the two demarcation and located it in falsifiable prediction; a theory earns its standing by forbidding something, by exposing a flank the world could, in principle, turn against it. This is correct as far as it goes. It does not go far enough, because a theory can issue falsifiable predictions at its periphery while keeping its core perfectly insulated, conceding small empirical skirmishes and never once risking the proposition that actually animates it.

Lakatos saw this and named the maneuver. A research program survives not by being right but by absorbing anomalies into a protective belt, reassigning blame to auxiliary assumptions so that the hard core is never the thing that takes the hit. A degenerating program does this reflexively; it meets every disconfirmation with a fresh epicycle and never with a funeral. The pathology is not that such a program makes no predictions; it makes many. The pathology is that none of them can reach the center. The center has been quietly placed beyond the range of its own evidence, and the placement is the whole trick. What presents itself as resilience is an architecture of self-exemption.

This suggests a different test than the one Popper proposed, and a harder one. The question is not whether a framework forbids something at its edges, but whether it can specify the configuration of its own central failure and then occupy that configuration when the evidence demands it. Most frameworks cannot, and the reason is structural rather than dishonest. A framework that could name, in advance and in its own vocabulary, the precise state in which its core claim would be dead has, in the naming, already half-killed it, and few are built to survive that operation. Instead, they are built not to perform it. They leave

the death-cell unlabeled, and an unlabeled cell is one you can always claim not to be standing in. To label it is to write, ahead of time, the obituary you may have to publish, to commit, before the evidence arrives, to the single outcome under which you agree to be finished. That is not a comfortable thing to ask of a theory, and it is not usually asked.

A theory, then, is not vindicated by what it confirms. It is vindicated by its capacity to close; by whether, when the world declines to cooperate, the framework can recognize the refusal as its own predicted ending rather than as someone else's measurement error. This is a strong requirement, and it is rarely met, because meeting it asks for a framework whose central thesis is, in effect, an account of how frameworks fail, turned without exemption on itself. I now have one in hand.

That framework is Recursive Coherence, and its central thesis is a theory of how symbolic systems fail. The thesis has a name, symbolic lock-in, and a precise content. A symbolic system enters lock-in when its internal coherence becomes self-referentially closed and decouples from the substrate it was built to track. It does not stop functioning; it functions too well at the wrong task. It amplifies the signals that confirm it, reclassifies the perturbations that would correct it as noise, and shifts its persistence logic from tracking reality to reproducing itself. The diagnostic signature is a single divergence, with confidence rising while the empirical anchor falls. A locked-in system grows more certain as it grows less attached, and that certainty is manufactured by the detachment, not in spite of it.

Stated this way, lock-in is a general account of the failure Lakatos described and Popper tried to fence out; the self-sealing core, the protective belt, or the program that survives by placing its center beyond its own evidence. But a general account of self-exemption has a property the theories Popper indicted did not. Marxism could explain away its predictive failures because nothing in Marxism flagged the explaining-away as a symptom. Psychoanalysis could absorb every outcome, confirmation and disconfirmation alike, because its theory of mind contained no clause that turned the absorption into a diagnosis. Their cores were silent about the maneuver that protected them, and silence is

what made protection possible. They could exempt themselves because they never wrote down what self-exemption looks like.

Recursive Coherence has written it down, and I take that consequence seriously. Its central claim is, in effect, a detector for the exact move a degenerating program makes to survive, and a detector cannot coherently decline to scan its own operator. If the thesis is true, then a symbolic system that amplifies its confirmations, prunes its disconfirmations, and persists by self-reproduction is failing; and a framework is a symbolic system; and a framework that did those things to protect its central claim would not be defending the claim but instantiating it, in the very failure mode the claim describes. The clause forbidding self-exemption was written into the foundation, which is either a discipline or an imprudence, depending on whether one intended to survive contact with one's own evidence.

So, the requirement for this framework is not an external standard imposed from the philosophy of science. It is an internal obligation. A framework whose central claim is failure-detection is not exempt from it. It is its own first case. The detector has to be turned on the instrument that built the detector, and the only honest way to do that is not in retrospect, once the verdict is in, but in advance, before the evidence arrives, while the outcome it would condemn is still a live possibility rather than a settled fact. That is what the framework did.

It did so in a paper whose title was already a confession; solving a validation problem that cannot be solved by validation. The problem was self-inflicted and unavoidable. Recursive Coherence denies that there is a static ground beneath the symbol stream, an oracle layer of mind-independent fact against which a measurement could be checked and stamped correct. That denial is load-bearing; it is the same refusal that makes lock-in diagnosable at all, since a symbolic system permitted to appoint itself its own ground would be immune to the charge of having lost contact with reality. However, a framework that forbids itself the oracle has also forbidden itself the ordinary way an empirical claim is discharged. It cannot validate its central measure against an authoritative external label, because importing such a label would falsify its ontology

before the first number was taken. Two honest moves remained, abandon the empirical ambition, or build a test that runs entirely inside the framework's own commitments.

I decided to build the test. In place of correspondence to an outside, it specified a set of internal signatures a self-consistent measurement would have to show jointly, that the measure's rankings persist under perturbation as the persistence criterion requires, that they converge across estimator families whose inductive biases genuinely disagree about what counts as structure, and that they separate a structural intervention from a merely interpretive one. Around these I laid an outcome matrix, a grid binding in advance what each configuration of results would be taken to mean, so that no verdict could be reinterpreted after the fact to the framework's advantage. The binding was the point. A test whose meaning is fixed only once the data are in is not a test; it is a negotiation.

One cell of that grid is the one that matters here. It is the configuration in which the estimators agree with one another, but the agreement is cut loose from persistence and from the substrate–symbol distinction; convergence without grounding, an internal consistency that tracks nothing outside itself. The framework did not leave this cell unlabeled. It named it, in its own vocabulary, as symbolic lock-in of the framework's own measurement operation, the instrument certifying itself by the concord of its own family. The line beside it is the single most self-aware moment in the corpus. *"The dictionary class agreeing with itself is not evidence that the substrate is compressible by dictionaries; it is evidence that dictionaries cohere with one another, which we already knew."*

This is the first true meta-moment, and it is worth pausing on what it actually is. A framework, denied the usual route to validation, sat down and wrote its own obituary; specifying the precise cause of death, the configuration of results under which its central instrument would be pronounced not wrong but *empty*, and then filed the document before the funeral, indeed before there was any body. This is not a hedge written for safety. It is a pre-committed self-indictment. A hedge widens your exits; it lets you claim, whatever happens, that you were never in the bad case. This narrows them to one. It commits the framework, in advance and in writing, to the single configuration under which it agrees to

be finished, and it hands the reader the criterion for checking whether that configuration obtained.

Two programs were tested. The first was Coherence Information Theory, the empirical arm of the framework, and its staked question was whether coherence is a measurable quantity beyond known statistical information; whether the weight the theory induces on a symbol stream carries structure the ordinary apparatus of information theory cannot already see. The test was the cross-domain audit the framework had specified. Derive the weight through several compression estimators, across domains of differing transparency, and ask whether their verdicts converge. They converged, but the convergence was the one the matrix had marked in advance as worthless, agreement within a family of estimators that, however different their surface mechanics, were all in the end statistical estimators, all measuring the same compressibility under different names. The convergence the test demanded was present; the grounding it demanded alongside, that the agreement track persistence, that it register the difference between substrate and symbol, was absent. That is the cell. The instrument that operationalized "coherence" was itself a compression-based estimator, and what such an instrument can detect is, by its own construction, statistical information. The measure had not found a quantity beyond statistics. It had found statistics wearing the framework's vocabulary, and the cross-domain agreement was the dictionary class agreeing with itself. The claim closed by reduction; coherence-as-a-measurable is the entropy rate under another name.

The second was steer-push, a separate program with a separate question; whether a physical difference can be a sign below life, whether the distinction between a difference that selects an outcome and one that merely causes it can be drawn in bare matter, with no organism, no observer, nothing that benefits. The program built the criterion, fixed a bar in advance, and ran a candidate chosen to be unimpeachable; a selection no free-energy description could tell apart from a mere persister, sitting below the threshold of anything that maintains itself. The candidate passed the operational gate. It was then examined for the one property that separates a sign from a recorded effect, the capacity to misrepresent, to stand for what is absent and so to be capable of being wrong, and it had none. What had

passed the gate was a reliable memory; a physical record that tracked its cause faithfully and could misrepresent nothing, because a sign needs an interpretant, something for which the record stands in, and the pre-life setting had been built precisely to exclude that. It met every condition but the last, and the last was the only one that mattered. The claim closed by reduction; what looks like aboutness in bare matter is reliably-coupled memory under another name.

Set side by side, the two closings are the same event in two domains, and the event is the cell. In each, a symbolic claim produced exactly the result the matrix had pre-labeled as empty; internal success, convergence in the first, a passed gate in the second, severed from the substrate the claim was supposed to reach. The measurable certified statistics and called them coherence; the sign certified memory and called it aboutness. Both had decoupled from the substrate they were built to track, and both had continued by reproducing their own vocabulary, which is precisely the definition the framework gives of symbolic lock-in. The framework had named, in advance, the configuration of its own death; its two most exposed claims walked into that configuration and occupied it. The framework did not survive its tests. It recognized itself in their failure and saw, in the collapse of each claim, not an anomaly to be explained away but a clean instance of the master failure mode it had spent its existence describing. Neither closing, however, was a refutation.

A refutation is a verdict the world hands down. You predicted something, the world produced something else, the claim is false. Neither closing was that, and the difference is the whole substance of this paper. To see it, one distinction has to be made sharply and then kept. A claim is *unconfirmed* when a confirming observation is possible but did not occur. You looked, and the thing was not there, though it could have been. A claim is *ill-posed* when no confirming observation is even constructible, when there exists no instrument that could register the "yes," because the instrument that would register it is built from the very thing the "yes" was supposed to exceed. The first is a fact about the world; it declined to cooperate. The second is a fact about the question; it cannot be put in

a form that admits the answer it wants. Both closings were of the second kind, and they were of the second kind by necessity, not by misfortune.

Take the measurable. The claim was that coherence carries structure beyond statistical information, but the only instrument that operationalizes coherence is a measure of compression, of predictive, statistical relevance, and by the Shannon–McMillan–Breiman theorem a universal compressor's rate converges, in the limit, to the entropy rate of the source. An instrument built from compression is, asymptotically, an estimator of statistics. To detect structure *beyond* statistics one would need an instrument that is not a statistical estimator and still measures coherence, and the framework has none, because it defines coherence as compression. The surplus was therefore not absent in the way an undetected thing is absent. It was undetectable in the way a thing is undetectable when the only available detector is made of the substance it was meant to exceed.

Take the sign. The claim was that a difference can be a sign in bare matter, but what distinguishes a sign from a faithful record is the capacity to misrepresent, and misrepresentation requires a standard of correctness, a fact about what the token *ought* to indicate, against which an off-target token counts as wrong rather than merely different. This is the disjunction problem, stated in matter; a token caused by X carries the information that X, or anything else that could have caused it, obtained, and a content that permissive can never be false. Only a function can be false, and a function requires something the indicating is *for*, a beneficiary. The below-life setting was built precisely to exclude the beneficiary, so the surplus that marks a sign could not be positive there. Not because no sign happened to turn up, but because the single property that would make something a sign has no instantiation in a setting stripped of the interpretant.

One wall, two faces. In each case the instrument that would have to register the surplus is the thing the surplus was meant to transcend; compression for the measurable, the interpretant for the sign. The transcendence was undetectable by construction, and the framework's own ontology had already implied as much, which is why the death-cell could

be named in advance. An instrument built from the substrate can only ever agree with its own family, and agreement with one's own family is precisely what the cell was.

These claims were not false. They were unaskable. The vocabulary that would have stated each surplus was borrowed, in the end, from the substrate the claim was trying to leave behind, but the asking began to draw something.

What it drew is a map with three regions and two borders, and the borders are the contribution. The regions are not new. A stratification of sign-ness runs from Peirce through Deacon, an iconic and indexical layer, where a difference resembles or is causally tied to what it indicates, and above it the symbolic, where reference is conventional, combinatorial, and detached from any single ground. The layering is theirs. What the two closings supply is not the layers but the lines between them, the two cuts the tradition had posited and never made with anything sharper than assertion.

Read in the framework's own terms, the map has a floor, a middle, and a ceiling. The floor is the physical substrate, and it is described without remainder by thermodynamics and information, by what persists and what is distinguishable, with no surplus quantity hiding beneath the statistics. The measurable close fixed that boundary; there is nothing above the entropy rate here to be found, because the only thing here is the entropy rate. The middle is the layer of sign and function, and it switches on only where there is something for an indication to be *for*, a system that can be served or harmed by what its records track, which is to say, at or above the threshold of life. The sign close fixed that boundary; misrepresentation, and with it aboutness, has no instantiation below the interpretant. And the ceiling is the symbol proper, the high-order compression that displaces, recombines, and refers across contexts; the human stratum, and, not incidentally, the only stratum in which symbolic lock-in can occur, because lock-in is a pathology of compressions that have forgotten what they compressed.

This is the error the framework had shared with the field it was reaching toward, and it was an error of a single word. Biosemiotics carries "sign" downward across all three regions, and the corpus, in its expansive moods, had done the same, calling a molecular event a selection, a bonding pattern an interpretation, as though one continuous semiosis

ran from the chemistry to the cortex. The closings cut that continuity exactly where it had to be cut. "Sign" does not reach the floor, because the floor has no surplus for it to name. "Sign" does not reach below the middle, because below life there is nothing the sign is for. The word had been doing three jobs at once, and the closings showed it was qualified for only one.

It is a humbler result than the one I set out for, and worth saying plainly. The framework went looking for a new continent, a quantity beyond information, a sign in bare matter, and returned, instead, with an accurate coastline. The closing did not empty the map. It drew the two lines the map had always been missing. A map, though, is only as good as the hand that drew it.

The hand is what survived. Strip away the two claims that closed and what is left is not a diminished quantity but a different kind of object, an instrument, not a measurement. The framework was never, in the end, a meter that returns a number called coherence; that meter was the part that broke. What did not break was the operation that broke it; the capacity to take a symbolic structure, hold it against the substrate it claims to track, and register the moment the two come apart. A reflexive lock-in detector is what the framework is, and the proof is not an argument but the preceding pages. It was turned on its own most cherished outputs, and it found them locked in. An instrument that can condemn its maker is at least an instrument.

Note where such an instrument has purchase and where it does not. Lock-in is a pathology of the symbol layer, of compressions high-order enough to forget what they compressed and stable enough to keep running once they have. It does not occur on the floor, where there is nothing but statistics to be tracked, nor in the middle, where function is too tightly coupled to its beneficiary to drift far before the coupling corrects it. It occurs at the ceiling, among human symbols, which is exactly the region the framework can still address. The discipline survives intact precisely because its object, the self-sealing symbolic structure, lives where the framework always had its best vision. The loss was a loss of reach, not of function. The meter failed at the bottom of the stack, and the diagnostic worked at the top, which is where the phenomena it cares about were all along.

There, the work that remains is real and unembarrassed by the closes. The framework reads psychiatric distress not as a set of intrinsic defects but as the signature of constrained option-space, depression as the collapse of futures in which action reliably improves outcomes, anxiety as the cost of monitoring a threat that admits no discharge, and it holds that these constraints are themselves observable, the breadth of available action, the load of what must be suppressed, and the latency before feedback returns. From this follows a wager that can actually fail; that large-scale changes in a population's constraints produce large-scale changes in its pathology, with the neurochemistry shifting after the constraint landscape and not before. That claim does not need a measurable coherence; it needs epidemiology, and epidemiology can return a verdict. The survivor is not a metaphysics in search of confirmation. It is a diagnostic with falsifiable work in front of it, all of it above the symbol line.

The right description of this framework has inverted over the course of its own falsification. Its value was never in what it could measure below the symbol. It is in what it can diagnose above it, including itself. That last clause is the one worth dwelling on, because it is not a fact about this framework. It is a property a framework can have, and naming it is the only result here that generalizes.

Call the property reflexive falsifiability, or the capacity of a framework to specify, in its own vocabulary, the configuration in which its central claim would be dead, and to occupy that configuration when the evidence arrives, recognizing the occupation not as a defeat administered from outside but as the arrival of its own prediction. It is rarer than ordinary falsifiability and it cuts deeper, because ordinary falsifiability lives at the periphery and this lives at the core. A framework can expose a hundred peripheral flanks and remain, at the center, untouchable. Reflexive falsifiability is the willingness to expose the center.

This property divides frameworks into two kinds, and the division is structural rather than moral. A degenerating program cannot name its own death-cell, because the naming would already be the death. To write down the configuration under which your core is empty is to half-empty it in the writing, and a program organized around the survival of its core cannot perform that act and remain what it is. So, it stays silent about its own ending,

and that silence is load-bearing. The other kind of framework cannot stay silent, because its central thesis is itself an account of how that silence works; it is a detector for the exact self-sealing the first kind depends on, pointed without exemption at its own operator. The first kind survives by never looking. The second survives only as long as it is right, and when it is wrong about something central it dies cleanly, because it has already told you where to look and what you would see.

This reframes the words we use to praise theories. A framework that has merely been confirmed has shown only that the world did not refuse it this time; it has not shown that it could recognize a refusal if one came. A framework that can name its death-cell and walk into it has shown the harder and stranger thing, that it has a death-cell at all and that it is the kind of object capable of dying. The capacity to die is not a weakness in an epistemic object; it is the precondition for the object being alive in the sense that matters. What cannot die is not robust. It is inert, and we already have a word for inert symbolic structure that goes on certifying itself.

There is a name in the corpus for the operation this paper has described. An earlier essay defines philosophy not as the construction of belief but as the recursive deconstruction of symbols that have fossilized into infrastructure, the pruning of the stagnant so that the conditions for future viability are preserved. What has happened here is that operation turned on its author. The framework took its own two most fossilized claims, the ones it most wanted to be true, and pruned them when they had stopped tracking anything but themselves. Its most convincing demonstration of the thesis that symbolic structures fail by self-sealing is not an ideology or an institution observed from a safe distance. It is the framework's own corpse, examined by the framework's own method. The theory's best evidence is itself.

A framework earns the name discipline, then, not when it is confirmed, but when it can name the shape of its own death and meet it. By that standard this one qualifies, and the qualification is itself a verdict the framework is passing on the framework, which is the place this paper has to end.

A word on what this is not. The results reported are theory-led, not a theorem; reflexive falsifiability has been exhibited in one framework, by one framework, and the claim that it cleanly divides all frameworks into the self-sealing and the self-pruning is a reading, offered as such, not a proof. What survives the two closings is narrow, a discipline that diagnoses lock-in in human symbolic structures, and a handful of psychiatric claims with empirical work still owed. That is the whole of the inheritance, and it is best not to inflate it; inflation is, after all, the failure this framework exists to name.

There is something faintly absurd in the position, and I will not pretend otherwise. A paper whose proudest result is that its own framework knew how to die is not the discovery its author set out to make. I went looking for a quantity beyond information and a sign in bare matter, and I have come back with an obituary and a coastline. The dignity available here is not the dignity of having been right. It is the smaller, stranger dignity of having built something that could tell you when it wasn't.

This paper is the latest pass of the same blade. To write it is to take the framework's two most cherished claims, hold them once more against the substrate they failed to track, and prune them in public; not to refute the framework but to leave, where the claims used to stand, the cleared ground the discipline was always for. Nothing here asks to be believed. It asks only to be useful after it is forgotten. Philosophy, in its highest recursive function, leaves no trace, only infrastructure.

References

James, B. (2026). coherence-information. Github. DOI 10.5281/zenodo.20399412. https://github.com/benjaminjamesgit/coherence-information

James, B. (2026). steer-push. Github. DOI 10.5281/zenodo.20947370. https://github.com/benjaminjamesgit/steer-push

James, B. (2026). Meta-Coherence: A Cross-Domain Empirical Specification for Validating Induced Coherence Weights Without Correspondence Truth. PhilPapers. https://philpapers.org/rec/JAMMAC-2

James, B. (2026). Recursive Coherence: Foundations for an Adaptive Philosophy of Science, Mind, and Society. PhilPapers. https://philpapers.org/rec/JAMRCF

James, B. (2025). Recursive Iconoclasm: Philosophy as Symbolic Disintegration into Coherence Infrastructure. PhilPapers. https://philpapers.org/rec/JAMRIP